


Exhibit P



Sponsorship Effects in Online Surveys

Charles Crabtree¹ · Holger L. Kern² · Matthew T. Pietryka² 

Published online: 22 May 2020

© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

Many academic surveys administered online include a banner along the top of the survey displaying the name or logo of the researcher's university. These banners may unintentionally influence respondents' answers since subtle contextual cues often have great impact on survey responses. Our study aims to determine whether these banners influence survey respondents' answers, that is, whether they induce sponsorship effects. For this purpose, we field three different studies on Amazon's MTurk where we randomly assign the sponsoring institution. Our outcome measures include survey questions about social conservatism, religious practices, group affect, and political knowledge. We find that respondents provide similar answers and exhibit similar levels of effort regardless of the apparent sponsor.

Keywords Sponsorship bias · Online surveys · Demand effects · Mechanical Turk · Social desirability

Introduction

Survey researchers often take great care in designing their instruments because they know that subtle changes to survey features can produce large changes in respondent answers. Indeed, a large literature suggests that survey responses can change with minor differences in stimuli such as interviewer characteristics (Cotter et al. 1982), question wording (Kinder and Sanders 1990), question order (Sigelman 1981), response-option order (Galesic et al. 2008), the survey sponsor (Bischooping and Schuman 1992), and even the activities the respondents were engaged in before participating in the survey (Druckman and Leeper 2012). With the rising popularity of online surveys for social science research (e.g., Clifford and Wendell 2016; Clifford

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11109-020-09620-7>) contains supplementary material, which is available to authorized users.

✉ Matthew T. Pietryka
matthew.pietryka@gmail.com

¹ Department of Government, Dartmouth College, Hanover, NH 03755, USA

² Department of Political Science, Florida State University, Tallahassee, FL 32306, USA

and Piston 2017; Krupnikov and Bauer 2014), scholars have rushed to evaluate the validity of samples recruited online (e.g., Clifford et al. 2015; Hauser and Schwarz 2016; Krupnikov and Levine 2014), but less attention has been paid to the effects of stimuli unique to online survey environments.

Online, self-administered surveys eliminate some of the potential environmentally-induced biases present in other survey modes, but they are *not* stimulus-free environments. Just as telephone interviewers provide cues about their backgrounds, attitudes, and expectations (Cotter et al. 1982), online surveys provide cues about the investigators that may influence the survey response. One such cue is the sponsoring university's name or logo, often displayed prominently in a banner at the top of the survey screen. Despite the prominence of these logos and the frequency of their use, little research has examined whether they influence responses. If they do, this influence would call into question a large body of published research because university names or logos are commonly displayed by default in the banners of surveys administered over popular survey platforms such as Qualtrics.¹

Existing research provides little clarity about the extent to which these banners might influence responses. On the one hand, many scholars emphasize the instability of survey responses in the face of even seemingly imperceptible stimuli (Burdein et al. 2006, 360). For instance, Connors, Krupnikov and Ryan (2019) investigate the effects of disclosing to subjects that survey responses will be shared in an *anonymized* public dataset. They demonstrate that such disclosures encourage some respondents to provide more 'don't know' responses and socially-desirable answers. On the other hand, several recent studies have found no discernible effect when altering small parts of the online survey experience (Mummolo and Peterson 2019; White et al. 2018). While several studies have examined the effects of sponsorship on survey recruitment (Edwards et al. 2014; Tourangeau et al. 2009; White et al. 2018), they provide little causal leverage for understanding whether the sponsor influences the content of survey responses.

To examine such sponsorship effects, we fielded three studies that experimentally varied the presence of banners displaying university logos. In each study, respondents were recruited from MTurk and then assigned to one of three groups. The first group saw no banner while the other two groups saw banners with one of two university logos. Study 1 used University of Notre Dame and Ohio State University logos and focused on questions about social conservatism and religiosity. Study 2 used UC Berkeley and Liberty University logos and focused on affect toward various social groups. Study 3 used Harvard University and Fitchburg State University logos and focused on political knowledge and effort. Across

¹ We contacted Qualtrics to determine how frequently universities placed their name or logo in their default survey header. Andrew Camp, a Qualtrics Product Specialist, responded through email on 2018-03-14. He confirmed that universities typically work with Qualtrics to create custom, branded survey formats, or themes, but licensing terms prevented him from disclosing descriptive statistics. He did attest that "universities do tend to prefer to set these custom themes as a default to encourage their users to use said theme." We also searched the EGAP registry for pre-analysis plans that included the terms 'qualtrics' ($n = 53$), MTurk' ($n = 31$), or 'mechanical turk' ($n = 66$). Of the 126 un-gated plans, there were 4 that showed screenshots of the actual survey instrument, 2 of which (50%) included a university logo.

all three studies, we find no evidence that banners systematically influence survey responses.

A Theory of Sponsorship Effects

As far back as the Hawthorne Works studies (Roethlisberger and Dickson 1939), scholars have documented numerous ways in which researchers might unwittingly influence the attitudes or behaviors of research participants. In this vast literature, the most direct analogue for our study of online sponsorship effects is the work that examines interviewer effects for surveys administered in person or by phone. This work suggests the interviewer may produce *participation* and *response* effects—who responds to surveys and *how* do they respond. Recent work has examined sponsorship effects on participation in online surveys (Edwards, Dillman and Smyth 2014; Tourangeau et al. 2009; White et al. 2018), but little research explores sponsorship-induced response effects in online settings.

Studies examining sponsorship effects on participation lack causal leverage for response effects. For instance, White et al. (2018) examine whether the purported investigator's gender, race, or ethnicity influence which MTurk workers respond to surveys. To do so, their experiment varied the name of the purported investigator in recruitment advertisements and consent forms. By randomly assigning the apparent gender and race of the investigator at recruitment, the design provides causal leverage for identifying participation effects. However, their study cannot identify the causal effects of sponsorship on survey responses because of the potential differences between the self-selected groups. Leeper and Thorson (2019) address this point by revealing purported sponsorship only after respondents were recruited through MTurk, randomly assigning an Aarhus University logo, a marketing firm logo, or no logo at all. They find only small differences in effort and socially desirable responses between the randomly assigned groups. Naturally, one might expect larger effects for universities with more salient reputations among US-based respondents.

Interviewer effects arise through at least two common routes, social desirability bias and satisficing (Holbrook et al. 2003). Under either route, the effects should be largest for sponsors with more established reputations. Social desirability bias occurs when respondents tell the interviewer what they think he or she wants to hear. For instance, respondents tend to report more support for gender equality when interviewed by a woman rather than a man (Huddy et al. 1997) and more support for racial equality when interviewed by an African American rather than a white person (Cotte et al. 1982). If online sponsors induce analogous effects, we should expect them for sponsors whose reputations send clear signals about desirable responses. Satisficing occurs when respondents fail to make a concerted effort to interpret a survey question, form a correct judgment, or report their judgment accurately. Just as more enthusiastic interviewers encourage greater effort from their respondents, surveys associated with elite universities may do the same in online settings.

Survey Experiments

With these expectations, we fielded three studies designed to study the effects of sponsorship cues conveyed by university logos. We recruited participants for each study using MTurk because it has become the most prevalent source of data for survey experiments published in the *American Political Science Review*, *American Journal of Political Science*, and *Journal of Politics* (Franco et al. 2017).²

In each of the three studies, the experimental manipulation consisted of the prominent display of one of several banners on survey webpages. Randomization was at the respondent level, with each banner having an equal probability of being assigned. Each survey question was displayed on its own page; the same banner was shown on each of these pages for a given respondent.³ Banners were sized to be equally tall and arranged so that they covered approximately the top third of the respondent's screen. All three studies were restricted to MTurkers with U.S.-based accounts; respondents whose IP addresses revealed that they participated from outside the 50 U.S. states were dropped from the sample. Question order and the ordering of response categories within questions were randomly assigned. The design and analysis plans for the three studies were pre-registered with the *Open Science Foundation*.⁴ Our true university affiliations were not revealed until respondents were debriefed, immediately after all responses had been collected.⁵ Participants in earlier studies were barred from participating in later studies. Participants were paid \$0.35.

Study 1

We began with a difficult test, examining sponsorship effects in relation to questions on which respondents often hold strong attitudes. Study 1 asked two questions that tap into social conservatism and two questions about religious practices. Using questions from the 2012 American National Election Study and the 2016 General Social Survey, respectively, we asked about views on abortion and birth control. Using question wording from the 2016 General Social Survey, we asked about the frequency of religious attendance and prayer.⁶

Our target sample size was about 2,000 respondents given the monetary resources available to us. Following our pre-registered analysis plan, we dropped respondents

² Researchers have noted a recent wave of low-quality, bot-like responses in surveys administered to MTurk workers. Most of these problematic responses come from duplicated GPS coordinates, which we drop before analysis. Using alternative indicators, we find only negligible evidence of bot-like behavior. For details, see Note 1 on page 25 of the online supplemental information (SI).

³ In many cases, researchers display multiple questions per page. By only displaying a single question on each page we maximize the chance that participants notice the banner while responding. This stronger-than-usual treatment implies that our inferences about the absence of sponsorship effects are likely to be conservative, making it less likely that sponsorship effects exist “in the wild.”

⁴ Pre-registration documents can be accessed at <https://osf.io/pfqtb/> (Study 1, fielded October 28–29, 2017), <https://osf.io/9sw2k> (Study 2, fielded December 18, 2017–January 2, 2018), and <https://osf.io/hpbqw> (Study 3, fielded March 5–April 11, 2018).

⁵ Online Figure A1 in the SI displays the consent form and debrief text used for all three studies.

⁶ See Online Tables A2–A5 for the question wording and response options.

whose latitude and longitude coordinates, as reported by Qualtrics, were not located within the 50 U.S. states. We also dropped respondents who did not agree to participate in the survey or who failed to complete the survey. Finally, we ordered submitted surveys by their submission times and dropped any surveys with latitude/longitude coordinates identical to previously submitted surveys.⁷ See Online Table A1 for details including attrition rates. Our realized sample size was $n = 1,646$.

We randomly assigned one of three banners to each respondent: a banner that was blank, a banner that displayed the *Ohio State University* logo, and a banner that displayed the *University of Notre Dame* logo. Notre Dame's prominence as a Catholic university lead us to expect that seeing the Notre Dame logo might induce respondents to report more frequent religious behaviors or give more conservative answers than seeing the Ohio State logo or no logo.⁸ We included the pure control (no logo) condition to preserve our ability to distinguish between effects stemming from seeing any logo and effects stemming from seeing a specific university logo.⁹

Following both classic (Fisher 1935) and modern (Freedman 2008a, b; Imbens and Rubin 2015; Young 2019) advice on the principled analysis of randomized trials, we rely on randomization inference (RI) throughout to test the *sharp null hypothesis* of absolutely no treatment effect for any respondent. Given the nominal level of our survey data we use the χ^2 test statistic. For each outcome, we approximate (using one million randomizations) exact RI p -values for each of the three two-way comparisons as well as a joint test, which consists of combining the χ^2 test statistics across the three two-way comparisons.¹⁰ Table 1 displays the results; Online Tables A2–A5 show the complete data.

As can be seen from Table 1, one out of the 12 p -values for the 12 two-way comparisons is below the pre-registered 5% statistical significance level but this result does not survive adjustment for multiple testing (last column). We also compute an omnibus p -value for all outcomes and treatments in our experiment (by combining all 12 χ^2 test statistics as for the joint tests); this p -value equals 0.58. Study 1 does not provide any evidence that participants' survey responses are affected by the purported sponsor of our survey.

⁷ We followed this procedure to prevent the same individual from participating in our survey several times using different MTurk IDs. This approach is more conservative than simply dropping repeated IP addresses since respondents might access the survey using several devices.

⁸ Ohio State and Notre Dame differ on other dimensions in addition to religious affiliation. More generally, any two universities will differ along many dimensions so additional work would be required to identify which dimensions drive differences in response patterns. Note, however, that this question is separate from the question we consider here, which is whether the bundle of traits represented by specific university logos affects survey responses.

⁹ An exact randomization inference balance test (approximated using one million randomizations) of the null hypothesis that the mean vote share of Donald Trump in the 2016 Presidential elections is the same across all three treatment groups yields a p -value of 0.18. For each subject, we looked up county-level vote shares based on the respondent's IP address. In the interest of full disclosure we note that balance tests were not pre-registered.

¹⁰ As in Young (2019), we stack the squared test statistics for each comparison and compute the test statistic $\beta^T \Omega \beta$, where β is the vector of stacked squared test statistics and Ω is the inverse of the covariance matrix of the stacked test statistics based on 1 million randomly chosen treatment assignments.

Table 1 RI p -values Study 1

	No logo vs. OSU	No logo vs. ND	OSU vs. ND	Joint
Abortion	0.27	0.31	0.43	0.50
Birth control	0.55	0.99	0.65	0.83
Religious attendance	0.90	0.96	0.97	0.99
Prayer	0.04	0.28	0.34	0.12

The table displays RI p -values for the four outcomes (rows) and three two-way comparisons (columns 1–3). The last column shows joint p -values that account for multiple testing within outcomes

Study 2

Given the null findings in the first study, we next examined sponsorship effects with questions particularly susceptible to response bias. Response biases are common with unreliable items because they provide multiple interpretations, encouraging respondents to choose the interpretation that places them in the most favorable light. We therefore examined responses to 101-point feeling thermometers, which typically exhibit low reliability because respondents cannot make fine-grained distinctions between ratings of, say, 30, 31, or 32 (Krosnick and Presser 2010). Despite their lack of reliability, feeling thermometers are widely used to measure affect toward groups and individuals. We ask respondents how favorable they feel toward each of six groups with clear ideological divisions (liberals, conservatives, Muslims, people on welfare, atheists, and feminists) and paired these questions with banners displaying universities with strong ideological reputations. Respondents were randomly assigned to banners showing no logo, the *Liberty University* logo, and the *UC Berkeley* logo.¹¹ Interviewer effects usually bias responses toward the interviewer's own views (West and Blom 2017). We therefore expect that on the conservative thermometer, individuals in the *Liberty University* condition will tend to offer more favorable ratings than those in the *no logo* condition. And respondents in the *no logo* condition should tend to be more favorable than those in the *UC Berkeley* condition. For the other thermometers, we expect the opposite ordering. For consistency, we therefore reverse-code the conservative thermometer so that all tests expect the highest scores for *UC Berkeley* and the lowest scores for *Liberty University*. Thus, our hypotheses here are one-sided, in contrast to Study 1 and Study 3.

Our target sample size was about 3,000 respondents and our realized sample size was $n = 2,339$ (see Online Table A1). Online Figure A3 shows the distributions for the six thermometer items by treatment group. Table 2 shows RI p -values. To make our inferences robust to outliers (Imbens and Rubin 2015, pp. 64–74) we use the difference in average ranks as our test statistic. We have three two-way comparisons (columns 1–3) for six outcomes (rows), yielding 18 comparisons in total. All tests are *one-sided*; cell entries showing * denote contrasts that failed to have the hypothesized sign. None of the p -values are below the 5% cutoff for statistical

¹¹ A balance test analogous to the one for Study 1 yields a p -value of 0.70.

Table 2 RI *p*-values Study 2

	No logo vs. UC Berkeley	No logo vs. Liberty U	UC Berkeley vs. Liberty U	Joint
Liberals	*	0.29	0.39	0.58
Conservatives	*	0.21	0.28	0.44
Muslims	0.12	*	0.43	0.27
People on welfare	0.23	*	*	0.47
Atheists	0.23	*	*	0.48
Feminists	0.33	*	*	0.63

The table displays RI *p*-values from one-sided tests for six outcomes (rows) and three two-way comparisons (columns 1–3). The last column shows joint *p*-values that account for multiple testing within outcomes

*Denotes tests in which the test statistic failed to have the hypothesized sign

significance. The last column shows joint *p*-values that account for multiple testing within outcomes. None is close to the statistical significance level we pre-registered. Finally, the omnibus *p*-value is 0.70; we fail to reject the sharp null hypothesis that any treatment has any effect on any outcome for any participant.

We also included a three-item self-monitoring battery in our survey (measured before exposure to the treatments) because high-self monitors are particularly prone to social desirability bias (Berinsky and Lavine 2012). As stated in our study's pre-registration, we investigated the one-sided hypothesis that treatment effects would be larger for respondents who score high on the scale than for respondents who score low. The distribution of the self-monitoring scale in our sample is shown in Online Figure A4; see the figure caption for details on the construction of the scale. Empirically, we compare treatment effects for respondents in the lower tertile to treatment effects for respondents in the upper tertile. Online Table A6 shows the resulting RI *p*-values. We find no evidence of treatment effect heterogeneity.¹²

Study 3

Past work has found interviewer effects for factual knowledge and respondent effort (West and Blom 2017). Some interviewers are better able to encourage respondents to exert the necessary effort to evaluate the survey question, summon their relevant beliefs, and choose the response option that best reflects these beliefs. In online surveys, more prestigious universities may encourage greater effort, but they may also

¹² An anonymous reviewer suggested to also examine treatment effect heterogeneity for the feeling thermometer scales for Muslims, people on welfare, atheists, and feminists using affective ideological polarization (Mason 2018), the difference in liberal and conservative thermometer scales. This analysis is only valid under the assumption that the difference in liberal and conservative thermometer scales is not affected by the treatments. We have included this exploratory analysis in Table A7 in the SI. We find no evidence of treatment effect heterogeneity.

encourage respondents to admit that they do not know the true answer to factual questions. Therefore, in Study 3 we investigated the effect of survey sponsorship on (i) respondents' willingness to admit not knowing the answers to relatively difficult political knowledge questions and (ii) the amount of effort respondents exerted when answering such questions. Moreover, we also investigated the relationship between survey sponsorship and measures of effort when respondents were asked to answer an open-ended essay question. We randomly assigned sponsor prestige with banners showing no logo, the *Harvard University* logo, and the *Fitchburg State University* logo.¹³

Our target sample size was about 3,000 respondents and our realized sample size was $n = 2,212$ (see Online Table A1). We measured political knowledge using six political knowledge items about (1) U.S. senators' term length; (2) the budget category on which the U.S. federal government currently spends the least; (3) the current Chief Justice of the United States; (4) the job currently held by Theresa May; (5) the current U.S. Attorney General; and (6) the current Secretary of State.¹⁴ We allow "Don't know" responses because their use provides a common source of interviewer effects (Krosnick and Presser 2010) and because they are substantively interesting to students of political knowledge (Jessee 2017).

We are interested in whether survey sponsorship affects the fraction of "Don't know" responses as well as respondents' effort as measured by the fraction of correct responses and response times. Online Tables A8–A14 show the distribution of answers to the political knowledge items by treatment group. Online Tables A15 and A16 show the distribution of the number of "Don't know" and correct responses by treatment group. Finally, Online Figure A5 shows the distribution of total response times by treatment group.

We also asked participants to write an open-ended essay.¹⁵ We coded three outcomes: the Flesch reading ease score, the number of characters, and the response time in seconds, all of which we interpret as measures of effort. Online Figures A6–A8 display the outcome distributions by treatment group.

Table 3 displays RI p -values. Test statistics are absolute differences in ranks (to make inferences robust to outliers) for all outcomes except % *Don't know* and % *Correct response*, for which we use absolute differences in means. None of the p -values is below 0.05. The omnibus p -value equals 0.84, providing no evidence against the sharp null hypothesis of no treatment effect for any outcome for any subject.

¹³ A balance test analogous to the ones for the other two studies yields a p -value of 0.86.

¹⁴ The Chief Justice, Theresa May, federal spending, and senate term length questions are variants of items from the 2012 ANES political knowledge battery, chosen because they were some of the most difficult items and because they could be readily extended to four response options plus "Don't know." The questions about the Secretary of State and Attorney General were taken from Ahler and Goggin (2017). On March 13, 2018, Donald Trump fired Secretary of State Rex Tillerson, making the question about the current Secretary of State meaningless. After the news broke, we deleted all 94 responses received after March 12, 2018 and replaced this item with a new question which asks which individual is currently a senator from Florida. Exact question wordings and response options are displayed in Online Tables A8–A14.

¹⁵ "In a paragraph, please tell us what being a good citizen means to you."

Table 3 RI *p*-values Study 3

	No logo vs. Harvard	No logo vs. FSU	FSU vs. Harvard	Joint
<i>Political knowledge items</i>				
% Don't know	0.40	0.80	0.56	0.69
% Correct response	0.69	0.94	0.64	0.88
Response time	0.83	0.27	0.19	0.37
<i>Essay</i>				
Number of characters	0.75	0.47	0.30	0.57
Flesch score	0.92	0.82	0.74	0.94
Response time	0.98	0.42	0.44	0.66

The table displays RI *p*-values for six outcomes (rows) and three two-way comparisons (columns 1–3). The last column shows joint *p*-values that account for multiple testing within outcomes

Table 4 Study 3: recall of banner by treatment group

	No logo	Harvard	Fitchburg State
I remember a logo, but I do not remember the organization	47	123	203
I do not remember a logo	626	168	144
Harvard University	2	437	6
Fitchburg State University	13	6	363
Brookings Institution	4	2	2
Cato Institute	0	2	2
Gallup	3	5	1
Massachusetts Institute of Technology (MIT)	5	7	15
Pew Research Center	5	1	5
Salem State University	2	3	10
% correct recall	89	58	48

By treatment group, the table displays the number of respondents falling into each response category when prompted with “A logo may have been displayed at the top of the last few pages. If so, do you recall what organization the logo represents?” at the end of the survey

It is worth noting that this null finding cannot simply be explained by participants not paying attention to the banners. Table 4 shows that 58% of participants in the Harvard group recall seeing the Harvard banner at the end of the survey. The corresponding number for respondents in the Fitchburg State University group is 48%.

Discussion

Our experiments were designed to explore the possibility of bias induced by the logos of universities that vary in terms of religiosity, ideology, and prestige on survey items measuring social conservatism, group affect, and political knowledge.

Scholars come from a variety of institutions, explore an ever growing set of survey items, and field their surveys to many different samples. Although we did not detect it in our three studies, sponsorship bias in survey responses may arise for other sponsors, items, samples, or some combination thereof.

Since we cannot possibly examine all of these possibilities, scholars should consider how their specific research design and research questions might make our results more or less applicable. For instance, our research focuses specifically on university sponsors, but many non-academic organizations also conduct online surveys.¹⁶ Likewise, social desirability bias requires respondents to anticipate the answers that would be most desirable. Such bias seems most likely when the topic of the survey aligns with prominent features of the sponsor. For example, the presence of a university logo may push respondents toward specific answers to questions about education policy or student loans. More generally, survey methodology will continue to evolve and some questions and survey interfaces may be more susceptible to sponsorship bias than those we examined. Therefore, replication and extension of our research to new topics and technologies will be essential going forward.

Scholars must also consider whether our results generalize to other subject pools. There is some evidence that MTurk workers may be somewhat more susceptible to social desirability effects (Behrend et al. 2011), perhaps because they tend to be more attentive than respondents from other online labor markets (Chandler et al. 2019). Such differences cannot explain our null results, however, since the just-cited studies suggest that subjects recruited through MTurk should be *more* sensitive to our experimental treatments than subjects drawn from other pools. Broadly speaking, across a variety of topics, results from MTurk samples differ little from those found in other samples (Bartneck et al. 2015; Coppock 2019; Mullinix et al. 2015). This work suggests but clearly does not guarantee that our results might generalize beyond survey respondents recruited through MTurk.¹⁷

We drew our samples exclusively from MTurk because it is the leading supplier of experimental samples (Franco et al. 2017). We offered pay that exceeded average rates for MTurk studies but conformed to the norm of offering at least minimum wage when scaled to an hourly rate (Andersen and Lau 2018). Had we offered greater compensation, perhaps we would have recruited subjects who would have been more attentive to the experimental treatments.¹⁸ The MTurk worker pool might also change over time as alternatives to MTurk become more prominent and popular. Alternatively, the ups and downs of the American economy might alter

¹⁶ We should note though that research varying whether the sponsor is purportedly academic or non-academic finds only minor differences in participation rates (Tourangeau et al. 2014) and no meaningful differences in response quality (Leeper and Thorson 2019).

¹⁷ Krupnikov and Levine (2014) suggest that MTurk samples sometimes yield results that differ from national samples in other important ways. The authors, however, issued a correction to their original findings (Krupnikov and Levine 2019), noting that many of the cross-sample differences they found lack statistical significance.

¹⁸ The literature casts doubt on that possibility, though. Wages in online labor markets such as MTurk seem to influence which workers are willing to participate but not the quality of the responses themselves (Mason and Watts 2009; Andersen and Lau 2018).

the composition of the MTurk worker pool. In any of these cases our experimental results might not necessarily generalize to future MTurk workers.

Given the innumerable dimensions of potential sponsorship effects, we recommend that scholars omit logos from their survey interfaces. Although our results suggest no harm in including them, they also reveal no benefits (such as increased effort). Since our results may not apply to all contexts, scholars can simply avoid the potential for bias by omitting institutional logos. Doing so will also enable researchers to include screenshots when submitting manuscripts for anonymous peer review. By including complete screenshots, researchers will help avoid the all-too-frequent mistake of failing to report all survey items and experimental treatments (Franco et al. 2015; Gerber et al. 2014).

Conclusion

Surveys conducted on MTurk routinely include university logos in banner images. Contrary to a large literature demonstrating that even seemingly minute alterations of the survey experience can induce changes in survey responses, we found no evidence that their prominent display affects respondents' reported religiosity and policy attitudes (Study 1), group affect (Study 2), or political knowledge and effort levels (Study 3). Since we cannot feasibly study a representative sample of universities or survey items, it is of course logically possible that sponsorship effects could exist for other universities or other survey items. Nonetheless, we examined the effect of logos for some of the most recognizable universities using items previously linked to various forms of response bias.¹⁹ If logo-induced sponsorship effects were widespread, we would expect some systematic differences between our treatment groups. Likewise, one might imagine larger effects for regionally-proximate respondents, alumni, or other subsets of individuals who would be particularly sensitive to such cues. Although possible, our results show that these groups are not sufficiently large portions of typical MTurk samples to lead to detectable sponsorship effects in our experiments (and, presumably, in other social science research using MTurk samples). Further, we show in Study 2 that even high self-monitors—a group prone to socially-desirable responses—fail to exhibit sensitivity to university logos. For these reasons, we believe that the threat of bias from such logos is small. We nonetheless recommend that researchers omit banners and other cues where possible since their inclusion offers no apparent benefit.

¹⁹ Our findings leave open the possibility that respondents are simply unfamiliar with universities such as Harvard, UC Berkeley, or Notre Dame and unaware of their reputations as elite, or liberal, or Catholic institutions. However, if these well-known universities indeed lack the cultural resonance required to produce sponsorship effects, it is hard to imagine that other universities would be more culturally resonant among U.S.-based survey respondents. Finally, even if respondents indeed lack knowledge of university reputations, our results would still suggest that including university logos does not bias survey responses.

Acknowledgements We thank Hans Hassell, Lauren Ratliff Santoro, and audiences at Florida State University and the 2018 Midwest Political Science Association meeting for useful feedback. All errors remain our own. The research reported here was approved by FSU's Human Subjects Committee (HSC 2017.22511 and 2018.25529)

Funding None.

Data Availability Data and replication code are available at <https://doi.org/10.7910/DVN/ZWAGXZ>.

Compliance with Ethical Standards

Conflicts of interest None.

References

- Ahler, D. J., Goggin, S. N. (2017). Assessing political knowledge: Problems and solutions in online surveys. SSRN Working paper.
- Andersen, D. J., & Lau, R. R. (2018). Pay rates and subject performance in social science experiments using crowdsourced online samples. *Journal of Experimental Political Science*, 5(3), 217–229.
- Bartneck, C., Duenser, A., Moltchanova, E., & Zawieska, K. (2015). Comparing the similarity of responses received from studies in amazons mechanical turk to studies conducted online and with direct recruitment. *PLoS ONE*, 10(4), e0121595.
- Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, 43(3), 800.
- Berinsky, A. J., & Lavine, H. (2012). Self-monitoring and political attitudes. In J. Aldrich & M. Kathleen (Eds.), *Improving public opinion surveys: Interdisciplinary innovation and the American National Election Studies*. Princeton: Princeton University Press.
- Bischoping, K., & Schuman, H. (1992). Pens and polls in Nicaragua: An analysis of the 1990 preelection surveys. *American Journal of Political Science*, 36(2), 331–350.
- Burdein, I., Lodge, M., & Taber, C. (2006). Experiments on the automaticity of political beliefs and attitudes. *Political Psychology*, 27(3), 359–371.
- Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., & Litman, L. (2019). Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior Research Methods*, 51(5), 2022–2038.
- Clifford, S., Jewell, R. M., & Waggoner, P. D. (2015). Are samples drawn from mechanical Turk valid for research on political ideology? *Research & Politics*, 2(4), 1–9.
- Clifford, S., & Piston, S. (2017). Explaining public support for counterproductive homelessness policy: The role of disgust. *Political Behavior*, 39(2), 503–525.
- Clifford, S., & Wendell, D. G. (2016). How disgust influences health purity attitudes. *Political Behavior*, 38(1), 155–178.
- Connors, E. C., Krupnikov, Y., & Ryan, J. B. (2019). How transparency affects survey responses. *Public Opinion Quarterly*, 83(S1), 185–209.
- Coppock, A. (2019). Generalizing from survey experiments conducted on mechanical turk: A replication approach. *Political Science Research and Methods*, 7(3), 613–628.
- Cotter, P. R., Cohen, J., & Coulter, P. B. (1982). Race-of-interviewer effects in telephone interviews. *Public Opinion Quarterly*, 46(2), 278–284.
- Druckman, J. N., & Leeper, T. J. (2012). Learning more from political communication experiments: Pre-treatment and its effects. *American Journal of Political Science*, 56(4), 875–896.
- Edwards, M. L., Dillman, D. A., & Smyth, J. D. (2014). An experimental test of the effects of survey sponsorship on internet and mail survey response. *Public Opinion Quarterly*, 78(3), 734–750.
- Fisher, R. A. (1935). *The design of experiments*. Edinburg: Oliver and Boyd.
- Franco, A., Malhotra, N., & Simonovits, G. (2015). Underreporting in political science survey experiments: Comparing questionnaires to published results. *Political Analysis*, 23(2), 306–312.

- Franco, A., Malhotra, N., Simonovits, G., & Zigerell, L. J. (2017). Developing standards for post-hoc weighting in population-based survey experiments. *Journal of Experimental Political Science*, 4(2), 161–172.
- Freedman, D. A. (2008a). On regression adjustments in experiments with several treatments. *Annals of Applied Statistics*, 2(1), 176–196.
- Freedman, D. A. (2008b). On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2), 180–193.
- Galesic, M., Tourangeau, R., Couper, M. P., & Conrad, F. G. (2008). Eye-tracking data: New insights on response order effects and other cognitive shortcuts in survey responding. *Public Opinion Quarterly*, 72(5), 892–913.
- Gerber, A., Arceneaux, K., Boudreau, C., Dowling, C., Hillygus, S., Palfrey, T., et al. (2014). Reporting guidelines for experimental research: A report from the experimental research section standards committee. *Journal of Experimental Political Science*, 1(1), 81–98.
- Hauser, D. J., & Schwarz, N. (2016). Attentive turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, 48(1), 400–407.
- Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly*, 67(1), 79–125.
- Huddy, L., Billig, J., Braccioldieta, J., Hoefler, L., Moynihan, P. J., & Pugliani, P. (1997). The effect of interviewer gender on the survey response. *Political Behavior*, 19(3), 197–220.
- Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge: Cambridge University Press.
- Jessee, S. A. (2017). “Don’t know” responses, personality, and the measurement of political knowledge. *Political Science Research and Methods*, 5(4), 711–731.
- Kinder, D. R., & Sanders, L. M. (1990). Mimicking political debate with survey questions: The case of white opinion on affirmative action for blacks. *Social Cognition*, 8(1), 73–103.
- Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research*. Bingley: Emerald Group Publishing.
- Krupnikov, Y., & Bauer, N. M. (2014). The relationship between campaign negativity, gender and campaign context. *Political Behavior*, 36(1), 167–188.
- Krupnikov, Y., & Levine, A. S. (2014). Cross-sample comparisons and external validity. *Journal of Experimental Political Science*, 1(1), 59–80.
- Krupnikov, Y., & Levine, A. S. (2019). Cross-sample comparisons and external validity: Corrigendum. *Journal of Experimental Political Science*. <https://doi.org/10.1017/XPS.2019.7>.
- Leeper, T. J., & Thorson, E. A. (2019). Should we worry about sponsorship-induced bias in online political science surveys? *Journal of Experimental Political Science*. <https://doi.org/10.7910/DVN/KKFS8Y>.
- Mason, L. (2018). Ideologues without issues: The polarizing consequences of ideological identities. *Public Opinion Quarterly*, 82(S1), 866–887.
- Mason, W. & Watts, D. J. (2009). Financial incentives and the “performance of crowds”. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*. HCOMP ’09 Paris, France: Association for Computing Machinery pp 77–85.
- Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science*, 2(2), 109–138.
- Mummolo, J., & Peterson, E. (2019). Demand effects in survey experiments: An empirical assessment. *American Political Science Review*, 113(2), 517–529.
- Roethlisberger, F. J., & Dickson, W. J. (1939). *Management and the Worker*. Cambridge: Harvard University Press.
- Sigelman, L. (1981). Question-order effects on presidential popularity. *Public Opinion Quarterly*, 45(2), 199–207.
- Tourangeau, R., Groves, R. M., Kennedy, C., & Yan, T. (2009). The presentation of a web survey, nonresponse and measurement error among members of web panel. *Journal of Official Statistics*, 25(3), 299–321.
- Tourangeau, R., Presser, S., & Sun, H. (2014). The impact of partisan sponsorship on political surveys. *Public Opinion Quarterly*, 78(2), 510–522.
- West, B. T., & Blom, A. G. (2017). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology*, 5(2), 175–211.

- White, A., Strezhnev, A., Lucas, C., Kruszewska, D., & Huff, C. (2018). Investigator characteristics and respondent behavior in online surveys. *Journal of Experimental Political Science*, 5(1), 56–67.
- Young, A. (2019). Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics*, 134(2), 557–598.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Political Behavior is a copyright of Springer, 2022. All Rights Reserved.